Coverage of Oncology Drug Indication Concepts and Compositional Semantics by SNOMED-CT®

Steven H. Brown MD[1, 2], Brent A. Bauer, MD[3], Dietlind L. Wahner-Roedler MD[3], Peter L. Elkin MD[3.] [1]Department of Veterans Affairs, [2]Vanderbilt University, [3]Mayo Clinic

Objective: To evaluate SNOMED-CT 's ability to represent simple and compositional concepts in FDA approved oncology drug indications.
Methods: Oncology drug indications were decomposed into single and compositional concepts. SNOMED-CT's coverage of single concepts and the semantics needed to create compositional concepts were evaluated using automated and manual techniques.
Results: SNOMED-CT covered 86.3% of single concepts present in oncology drug indications; 11.3% of indications were covered completely. Coverage was best for concepts describing diseases, anatomy, and patient characteristics. Medications accounted for 50.5% of missing concepts. Excluding drug names, 45.2% of indications were completely represented. SNOMED-CT's semantics completely represented 60.1% of compositional expressions.
Conclusions: SNOMED-CT's overall coverage of the concepts in oncology drug indications was good. Improvements or alternatives are needed for medications and semantics.

Introduction

In the past five years a number of papers detailing desirable characteristics of terminologies have been published. In 1998, Chute documented 11 characteristics that terminologies should have or evolve to have in order to meet important needs of health care[1]. Cimino's [2] work from the same year described 12 "desiderata" synthesized from the literature of medical vocabulary research. ASTM E 2087-00, published in 2000, enumerated over 50 quality indicators for controlled health vocabularies[3]. ISO TS17117[4] carries forward the ideas in ASTM 2087 as an international technical specification. Two additional publications[5, 6] advance our understanding of terminology quality indicators even further. While the guideline authors may disagree on certain fine points, the importance of content coverage is universally acknowledged. In our experience, the importance of content coverage is understood and accepted by technical and non-technical audiences alike. "Content, content, content" [2] delivers the message succinctly.

Content coverage studies are not new to the literature. For example, in 1977 Lowery et al examined ICD, SNOMED, and the Cardiff system for coding congenital malformations and genetic syndromes[7]. A number of subsequent content coverage studies further evaluated the SNOMED family of terminologies[8-14].

SNOMED-CT is a reference terminology created from the combination of SNOMED-RT and the National Health Service's Clinical Terms version 3[15]. According to the July 2002 fact sheet, SNOMED-CT contained 333,000 concepts and approximately 1,000,000 "is a" semantic relationships. SNOMED-CT supports the composition of new terms through the combination of existing concepts. A national license for SNOMED-CT was being negotiated by the NLM at the time this manuscript was written. If this license agreement comes to pass, SNOMED-CT could become a defacto national standard. Thus, understanding the content coverage of SNOMED-CT is of particular importance at this time.

Compositionality has been proposed and successfully demonstrated as an approach to improve content coverage[16-18]. For example, post coordinated composition of UMLS concepts to represent problem statements has performed significantly better than UMLS concepts alone[19]. The linkage of two or more concepts is typically achieved using a formal semantic that details the concepts' relationship. For example, the concepts "enalapril" and "angiotensin converting enzyme inhibition" could be joined by the semantic relationship "has mechanism of action." Post coordinating a terminology's concepts via its semantics suggests another type of study: the content coverage of the linking semantics. We believe semantics are an important part of compositional terminologies. Others agree. For instance, Bakken evaluated SNOMED-CT's semantics in a study of nursing diagnoses[20].

In the current study, we evaluate SNOMED-CT's ability to represent the content of a set of FDA approved oncology drug indications and perform a preliminary analysis of its semantics

Methods

Approved oncology drug indications (table 1) were downloaded from the FDA Oncology Tools website[21]. SNOMED-CT version 1.0 from the College of American Pathologists was employed. All downloaded indications were manually broken into single concepts and compositional concepts. Our method identified the shortest medically sensible compositional concepts within the indication. Expressions composed of two concepts (e.g. oral + capsule) were identified whenever possible. A second author verified each proposed compositional expression. Examples of single and compositional concepts identified within indications are given in table 1. Each single or compositional concept was categorized as relating to treatments, diseases, patients, medications, anatomy, or other. Only concepts that mentioned a specific medication were classified as medication related. Concepts referring to broad classes of medications were classified as treatment related. Descriptive statistics and tables documenting the most commonly occurring single and compositional expressions are presented in the results section.

SNOMED-CT's content coverage of the identified single concepts was measured in two phases. In the first phase, automated concept identification tools available in our lab[19, 22] were applied to each indication. The output was an XML file containing the original indications and all mapped SNOMED concepts. Each indication concept to SNOMED concept mapping was manually reviewed for correctness. The indication concepts that were not mapped properly via the algorithmic approach were manually reviewed using the Mayo Vocabulary Server and Browser tool loaded with SNOMED-CT. In this manner, single concepts were determined to be present or absent.

SNOMED-CT's coverage of the semantics needed to form compositional expressions was evaluated by manual modeling. The single 'best' fitting semantic relation was used to link concepts forming each compositional expression. The adequacy of each semantic's representation of the meaning of the compositional expression was judged by consensus of two reviewers to be 1) complete, 2) partial, or 3) inadequate.

Results

The FDA website contained 115 indications for 68 unique drugs. We identified 1527 concepts in the 115 indications. The mean number of concepts per indication was 13.3 (95% CI 12.0 – 14.2) with a range from 3 to 48. Table 1 shows two representative indications and the concepts identified within them.

The ten most commonly found single concepts and their frequency of occurrence are: "Patients" (56), "Treatment" (44), "Cancer" (44), "Therapy" (34), "Combination" (34), "Metastatic" (25), "Breast Cancer" (24), "Advanced" (22), "Cell" (21), and "Approved" (17). The most represented category of concepts was treatment related (45.5%). The distribution of classes (treatment, medication, patient, anatomy, disease, other) for the single concept expressions is shown in table 2.

We identified 303 occurrences of 201 unique compositional concepts within the 115 indications. The most frequent compositional concept was used 11 times and 46 compositional concepts were used more than once. The ten most commonly found compositional concepts and their frequency of occurrence are: "First-line Treatment" (17), "Approved Chemotherapeutic Agents" (9), "Single Agent" (9), "Postmenopausal Women" (7), "Palliative Treatment" (7), "Adjuvant Therapy" (4), "Advanced Breast Cancer" (4), "Initial Chemotherapy" (4), "Disease Progression" (4), and "Hormone Receptor Positive"(4). Of these 303 compositional expressions, 290 were composed of two 'atomic' concepts and 13 were created using 3 'atomic' concepts. Table 3 shows the distribution of classes (treatment, medication, patient, anatomy, disease, other) for the compositional expressions.

SNOMED-CT covered 1317 of 1527 (86.3%) single concepts present in the oncology drug indications. Thirteen indications had all of their concepts covered (11.3%). Medication name was the most common category of missing concepts (50.5% of missing concepts). Excluding the 340 occurrences of medication names, SNOMED-CT covered 1083 of 1187 concepts (91.2%) and 52 of 115 (45.2%) indications completely. Table 2 details the frequency of missing concepts by category. Table 2 also provides the ratio between missing concepts and overall concept use for each category. Ratios less than one indicate good content coverage for the category.

Table 1. Example indications and concepts identified.

Drug & Trade Name	Indication	Concepts	Compositional Concepts
Alitretinoin Panretin	Topical treatment of cutaneous lesions in patients with AIDS-related Kaposi's sarcoma.	[Topical] [Treatment] [Cutaneous] [Lesions] [Patients] [AIDS] [Related] [Kaposi’s Sarcoma]	[Topical treatment] [Cutaneous lesions] [AIDS-related Kaposi’s Sarcoma]
Anastrozole Arimidex	For the adjuvant treatment of postmenopausal women with hormone receptor positive early breast cancer	[Adjuvant] [Treatment] [Postmenopausal] [Women] [Hormone] [Receptor] [Positive] [Early] [Breast Cancer]	[Adjuvant treatment] [Postmenopausal women] [Hormone receptor] [Early breast cancer]

Table 2. Comparison of single concept categories found within oncology drug indications and SNOMED’s coverage of them. A ratio of missing % to overall % less than 1 indicates relatively good coverage of that concept category.

Single Concept Category	Overall Representation	Overall %	Missing Concepts %	Missing/Overall Ratio	Missing Concepts	Unique Missing
Treatment	695	45.5	40.5	0.89	85	50
Medication	340	22.3	50.5	2.26	106	64
Disease	298	19.5	1.4	0.07	3	3
Patient	106	6.9	0.5	0.07	1	1
Anatomy	52	3.4	0.5	0.15	1	1
Other	36	2.4	6.7	2.79	14	11

Table 3. Comparison of compositional concept categories found within oncology drug indications and SNOMED’s linking semantics coverage of them.

Composition Category	Overall Representation	Overall %	Complete Meaning	Complete %	Partial Meaning	Partial %	Missed Meaning	Missed %
Treatment	172	54.4	103	59.9	50	29.1	19	11.0
Medication	12	3.8	3	25.0	5	41.7	4	33.3
Disease	107	33.9	66	61.7	38	35.5	3	2.8
Patient	14	4.4	10	71.4	4	28.6	0	0.0
Anatomy	1	0.3	1	100.0	0	0.0	0	0.0
Other	10	3.2	7	70.0	2	20.0	1	10.0

Table 4. Commonly used linking semantics.
.

Linking Semantic	Overall Rank	Overall Use	Complete Capture Rank	Complete Capture Use	Complete Capture %
Course	1	41	2 (tie)	27	70.7
Has intent	2	39	1	29	69.2
Occurrence	3	38	2 (tie)	27	71.1
Priority	4	27	4	26	96.3
Subject of information	5 (tie)	19	17 (tie for last)	1	5.3
Pathologic process	5 (tie)	19	17 (tie for last)	1	5.3
Finding site	10	13	5	13	100

We extracted 37 semantic relationships from SNOMED -CT and used 25 of them to model each of the 303 ‘simple’ compositional concepts present in the oncology drug indications. We were able to identify a candidate linking semantic for 289 of 316 (91.5%) pairs of atomic concepts (13 concepts required a second linking semantic). Semantics judged to partially capture the needed meanings numbered 99 (31.3%), and 190 (60.1%) were judged to completely capture the needed meaning. We were unable to identify a linking semantic that even partially captured the necessary meaning in 27 cases (8.5%). Table 3 shows the acceptability of the linking semantic broken out by category (treatment, medication, patient, anatomy, disease other). The first two columns of table 4 shows the rank and frequency of use of the most commonly employed linking semantics (n = 289). Columns 3 and 4 of table 4 show the rank and frequency of use of the most commonly employed semantics that were judged to have completely captured the needed meaning (n = 190).

We created a candidate semantic relation for each of the 126 instances in which the available SNOMED -CT semantics failed to completely capture the needed meaning. Of the 34 newly created semantics, the most commonly used was “has disease extent” (23 times). Eighteen of the new semantics were used more than once. Other commonly used new semantic relations include: “has status”, “has cardinality”, “has treatment outcome” and “has goal”.

Discussion

SNOMED -CT’s concept coverage of single concepts present in oncology drug indications was 86.3%. Excluding drug names, SNOMED - CT’s coverage improved to 91.2% of single concepts. SNOMED-CT performed relatively well in representing the source concepts relating to diseases, patients and anatomy and relatively poorly representing concepts relating to medications. These results reflect the traditional strengths of the SNOMED family of terminologies. Our result of SNOMED-CT’s single concept content coverage is similar to the 93% figure found by McClay et al[14] for SNOMED -CT’s coverage of primary reasons for visits to the emergency room.

Post-coordination is a recognized approach to improving content coverage. Formal semantics to link concepts are a workable approach to post-coordination. This approach adds flexibility and expressiveness to a terminology but requires that it include semantics with appropriate meaning. We found SNOMED-CT’s linking semantics to be completely adequate in 60.1% of cases. This is in contrast to the findings of Bakken et al[20], which found SNOMED-CT to include 8 of 9 (88.9%) of the semantics present in the ISO and CEN models of nursing diagnoses.

Oncology drug indications are complex expressions containing concepts related to drugs, diseases, treatment protocols, patient characteristics, and anatomy. This complexity explains why only 11.3% of indications were completely described by SNOMED-CT despite, its coverage of 86.3% of single concepts. Excluding drug names SNOMED-CT represented completely 45.2% of the indications. This complexity may also explain some of the differences between SNOMED-CT’s coverage of oncology drug indications and reasons for emergency room visits.

Other content coverage studies of SNOMED have published results ranging from 30% of health and functional status terms [13] to 70% of concepts from a mix of domains[11]. Direct comparison to the current study results is difficult for two reasons. First, most previous content studies used previous versions of SNOMED, not SNOMED-CT. Second, different sources were used to create the test concepts. Although using different sources of concepts makes direct comparisons of study results difficult, it is a necessity with some virtue. Terminology developers could easily incorporate a single test set of concepts. Additionally, different sources of test concepts better reflect the diversity of potential uses of controlled terminology.

We caution readers that this content coverage evaluation, like many others, includes subjective elements. We only examined indications for oncology drugs, not drugs approved for other purposes. In addition, we only examined one version of SNOMED -CT. We fully expect that it will improve and evolve over time. Given these caveats, it is our hope that the current study will contribute to an emerging understanding of SNOMED -CT. We applaud the developers for their considerable effort, and encourage them to continue on the course of ongoing improvement.

Supported by NLM grant LM06918-A101. The authors thank Russ Hamm and Constantin Mitocaru for their technical assistance.

References

1. Chute CG, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the United States. J Am Med Inform Assoc 1998;5(6):503-10.
2. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403.
3. ASTM-E31. E2087-00 Standard specification for quality indicators for controlled health vocabularies. Conshohocken, PA: American Society for Testing and Materials; 2000.
4. ISO/TC 215 Health informatics. ISO 17117 Health informatics — Controlled health terminology — Structure and high-level indicators. ISO Technical Specification 2002;2002-02-15:32.
5. Elkin PL, Brown SH, Chute CG. Guideline for health informatics: controlled health vocabularies--vocabulary structure and high-level indicators. Medinfo 2001;10(Pt 1):191-5.
6. Elkin PL, Brown SH, Carter J, Bauer BA, Wahner-Roedler D, Bergstrom L, et al. Guideline and quality indicators for development, purchase and use of controlled health vocabularies. Int J Med Inf 2002;68(1-3):175-86.
7. Lowry RB, Rocheleau J, Keillor L. Comparison of existing classifications for coding congenital malformation and genetic syndromes. Birth Defects Orig Artic Ser 1977;13(3A):53-9.
8. Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. Proc Annu Symp Comput Appl Med Care 1994:201-5.
9. Henry SB, Holzemer WL, Reilly CA, Campbell KE. Terms used by nurses to describe patient problems: can SNOMED III represent nursing concepts in the patient record? J Am Med Inform Assoc 1994;1(1):61-74.
10. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. J Am Med Inform Assoc 1996;3(3):224-33.
11. Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. J Am Med Inform Assoc 1997;4(3):238-51.
12. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc 1997;4(6):484-500.
13. Ruggieri AP, Elkin P, Chute CG. Representation by standard terminologies of health status concepts contained in two health status assessment instruments used in rheumatic disease management. Proc AMIA Symp 2000:734-8.
14. McClay JC, Campbell J. Improved Coding Of The Primary Reason For Visit To The Emergency Department Using SNOMED. Proc AMIA Symp 2002:499-503.
15. Wang AY, Sable JH, Spackman KA. The SNOMED Clinical Terms Development Process: Refinement and Analysis of Content. Proc AMIA Symp 2002:845-9.
16. Rassinoux AM, Miller RA, Baud RH, Scherrer JR. Compositional and enumerative designs for medical language representation. Proc AMIA Annu Fall Symp 1997:620-4.
17. Elkin PL, Tuttle M, Keck K, Campbell K, Atkin G, Chute CG. The role of compositionality in standardized problem list generation. Medinfo 1998;9(Pt 1):660-4.
18. Brown PJ, Price C. Semantic based concept differential retrieval & equivalence detection in clinical terms version 3 (Read Codes). Proc AMIA Symp 1999:27-31.
19. Elkin PL, Bailey KR, Chute CG. A randomized controlled trial of automated term composition. Proc AMIA Symp 1998:765-9.
20. Bakken S, Warren JJ, Lundberg C, Casey A, Correia C, Konicek D, et al. An evaluation of the usefulness of two terminology models for integrating nursing diagnosis concepts into SNOMED Clinical Terms((R)). Int J Med Inf 2002;68(1-3):71-7.
21. Food and Drug Administration. Listing of Approved Oncology Drugs with Approved Indications. 2003; Accessed: 2/6/03.; Available: www.fda.gov/cder/cancer/druglistframe.htm.
22. Elkin PL, Bailey KR, Ogren PV, Bauer BA, Chute CG. A randomized double-blind controlled trial of automated term dissection. Proc AMIA Symp 1999:62-6.